Methodenbeitrag: Möglichkeiten der Textdigitalisierung

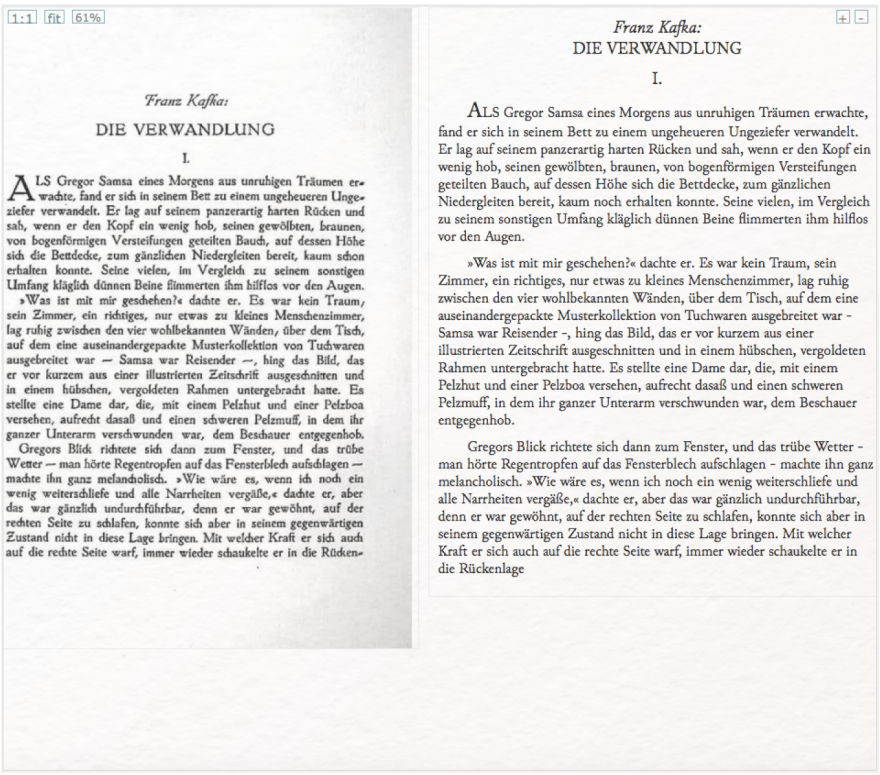
Jan Horstmann [1]

1. Universität Münster

Thema:	Textdigitalisierung und Edition	DOI:	10.48694/fortext.3741
Jahrgang:	1	Ausgabe:	3
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2018-01-24 auf fortext.net
Lizenz:	CC BY SA		open ∂ access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition



Franz Kafka:
DIE VERWANDLUNG

I.

ALS Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt. Er lag auf seinem panzerartig harten Rücken und sah, wenn er den Kopf ein wenig hob, seinen gewölbten, braunen, von bogenförmigen Versteifungen geteilten Bauch, auf dessen Höhe sich die Bettdecke, zum gänzlichen Niedergleiten bereit, kaum noch erhalten konnte. Seine vielen, im Vergleich zu seinem sonstigen Umfang kläglich dünnen Beine flimmerten ihm hilflos vor den Augen.

»Was ist mit mir geschehen?« dachte er. Es war kein Traum, sein Zimmer, ein richtiges, nur etwas zu kleines Menschenzimmer, lag ruhig zwischen den vier wohlbekannten Wänden, über dem Tisch, auf dem eine auseinandergepackte Musterkollektion von Tuchwaren ausgebreitet war – Samsa war Reisender –, hing das Bild, das er vor kurzem aus einer illustrierten Zeitschrift ausgeschnitten und in einem hübschen, vergoldeten Rahmen untergebracht hatte. Es stellte eine Dame dar, die, mit einem Pelzhut und einer Pelzboa versehen, aufrecht dasaß und einen schweren Pelzmuff, in dem ihr ganzer Unterarm verschwunden war, dem Beschauer entgegenhob.

Gregors Blick richtete sich dann zum Fenster, und das trübe Wetter – man hörte Regentropfen auf das Fensterblech aufschlagen – machte ihn ganz melancholisch. »Wie wäre es, wenn ich noch ein wenig weiterschliefe und alle Narrheiten vergäße,« dachte er, aber das war gänzlich undurchführbar, denn er war gewöhnt, auf der rechten Seite zu schlafen, konnte sich aber in seinem gegenwärtigen Zustand nicht in diese Lage bringen. Mit welcher Kraft er sich auch auf die rechte Seite warf, immer wieder schaukelte er in die Rückenlage

Kafkas Verwandlung in analoger und digitalisierter Form

Textdigitalisierung bezeichnet den Prozess der Umwandlung eines gedruckten oder handschriftlichen Textes in einen maschinell lesbaren elektronischen Text. Je nach Beschaffenheit des Ausgangstextes kommen in diesem Prozess der Texterfassung bzw. Transkription mehrere potentielle Bearbeitungsschritte in Frage – automatisierte (optical character recognition (OCR): optische Zeichenerkennung) (vgl. OCR) wie manuelle (keying) (vgl. Keying).

2. Anwendungsbeispiel

Angenommen, Sie wollen das Prosaoeuvre von Ingeborg Bachmann (vgl. Korpus) erforschen und sich dabei digital unterstützen lassen. Das Problem ist, dass Bachmanns Texte nicht gemeinfrei und damit nicht als fertige Textdigitalisate zugänglich sind. Sie haben nun die Möglichkeit, die gedruckten Texte selbst in digital lesbare Textdateien umzuwandeln, um anschließend die Methoden und Tools der digitalen Textanalyse nutzen zu können. Zu welchen Problemen es bei einer automatisierten Texterfassung kommen kann, wird anhand eines Ausschnittes aus Bachmanns Erzählung *Das Gebell* (1972) unter Punkt 5 dieses Beitrags veranschaulicht.

Texte, die urheberrechtlich geschützt sind, dürfen jedoch nur für den privaten bzw. eigenen wissenschaftlichen Gebrauch digitalisiert und nicht veröffentlicht oder vervielfältigt werden.

3. Literaturwissenschaftliche Tradition

Die Digitalisierung von Texten führt mehrere Traditionslinien fort: (1) die Editionsphilologie und Textkritik, (2) die Paläografie und auch (3) das Setzen von Manuskripten seit der Erfindung des Buchdrucks.

1) Mit der Erfindung der Schrift vor über 5000 Jahren wurde eine Möglichkeit geschaffen, die auch als Leitbild der Textdigitalisierung verstanden werden kann: „die konsistente Bewahrung der sprachlichen Äußerung durch sie repräsentierende Zeichen“ (Nutt-Kofoth 2007, 1). Möchte man einen Text elektronisch lesbar machen, sollten einige Maximen der Editionswissenschaft, die sich mit dieser konsistenten Bewahrung beschäftigt, beachtet werden. Dies gilt insbesondere für Texte, die selbst aus historisch-kritischen Ausgaben stammen, denn hier ist eine korrekte Texterfassung die Basis für jede wissenschaftlich tragbare Weiterverarbeitung und auch in der elektronischen Version sollten einzelne Elemente der Edition (Sproll 2007, 178) differenziert werden können. Historisch-kritische Ausgaben, kritische Ausgaben, Studienausgaben oder Leseausgaben stellen in der Editionswissenschaft unterschiedliche Ansprüche an die Authentizität der Textwiedergabe. In vielen Ausgaben der einzelnen Editionsvarianten (außer der (historisch-)kritischen) greifen die jeweiligen Herausgeber*innen normalisierend in Orthografie und Interpunktion ein (Nutt-Kofoth 2007, 6f.). Was in der Editionsphilologie ein „diplomatischer Abdruck“ (Grubmüller und Weimar 1997, 414) genannt wird, sollte jedoch auch das Ziel der digitalen Erfassung eines gedruckten Textes sein: die „urkundlich genaue, zeichengetreue Wiedergabe eines Textes, d. h. unter Bewahrung aller Besonderheiten und auch Uneinheitlichkeiten von Orthografie und Interpunktion“ (ebd.). Historische Orthografie korrekt abzubilden, ist für die automatische Texterkennung (OCR), die ihre Ergebnisse auf Grundlage aktueller Rechtschreibung korrigiert, eine Herausforderung, jedoch ein wichtiger Orientierungspunkt, denn sie „ist nichts Äußerliches, Akzidentelles, sondern als ein Teil der historischen Form bedeutungskonstituierendes Merkmal“ (Kraft 2001, 72). Auch die Wahrung des Lautstandes ist hierbei ein wichtiger Grund der originalgetreuen Wiedergabe (Kraft 2001, 85). Der Umgang mit dem Original statt einem Faksimile ist eine Maxime der textkritischen Arbeit, denn „[b]ereits durch eine Verkleinerung oder eine Vergrößerung des Originals tritt ein ‚Informationsverlust‘ ein“ (Kraft 2001, 63). Der im ersten Schritt der Textdigitalisierung erstellte Scan (s. u.) muss demnach dem Original so ähnlich wie möglich sehen, um einen solchen Informationsverlust zu vermeiden. Die Prinzipien der Textkritik, die sich im engeren Sinne mit der Herstellung eines Archetypus anhand überlieferter Textzeugen eines verlorengegangenen Textes beschäftigt, in den neueren Philologien jedoch in der Editionstechnik aufgeht, finden im übertragenen Sinne ebenfalls eine Fortführung in der Textdigitalisierung. Recensio (mustern), examinatio (prüfen) und emendatio (verbessern) (Kocher 2007, 761) sind auf den Prozess der Vergleichsarbeit von gedrucktem und digitalisiertem Text übertragbar, wobei hier freilich der gedruckte Text als das Original vorliegt und nicht erschlossen werden muss. Dennoch müssen beispielsweise auch hier „Druck- und Schreibfehler sowie Schäden des Textträgers“ (Kocher 2007, 762) als solche identifiziert und ein Umgang mit ihnen gefunden werden.
2) Auch die Digitalisierung von Handschriften geht zurück auf eine literaturwissenschaftliche Tradition im weiteren Sinne bzw. eine ihrer Hilfswissenschaften, nämlich die Paläografie, in der es „ganz allgemein um das Lesenkönnen der alten Schriften […] aber auch die Physiologie und Psychologie des Schreibens“ (Rohr 2015, 125) geht (Schneider 2003, 1). Wesentliche Kategorien sind hier u. a. die Differenzierung unterschiedlicher Schreib- und Beschreibstoffe sowie die Identifikation von Abkürzungen und Sonderzeichen, des Duktus, Ligaturen, des Schriftwinkels oder der jeweils gebrauchten Schriftart (z. B. Druck- oder Schreibschrift, vereinfachte Ausgangsschrift, Majuskel- oder Minuskelschrift etc.), denn „verschiedene Schriftarten entstehen für unterschiedliche Zwecke, lassen sich in ihren Entwicklungsphasen und Blütezeiten nachvollziehen und werden schließlich von neuen Erscheinungen unterlaufen und letztlich abgelöst“ (Schneider 2014, 13).
3) Schließlich finden sich auch Veränderungen, die im Zuge des Buchdrucks die Umwandlung von Manuskripten in einen Drucksatz betrafen, in der heutigen Textdigitalisierung wieder. Beim Setzen geht es darum, das Manuskript buchstabengetreu wiederzugeben und häufig wurde dies durch einen Setzer und nicht den Autor selbst ausgeführt. „Eine der großen editorischen Schwierigkeiten bezüglich der Literatur der Frühen Neuzeit, der ersten Jahrhunderte nach Gutenberg bis zum Barock, besteht in der Klärung der Frage, […] wie sich der Text durch vielfache Wiederabdrucke verschlechtert hat“ (Nutt-Kofoth 2007, 11). Diese Frage sollte auch an ein Textdigitalisat gestellt werden.

4. Diskussion

Die automatische Texterfassung mittels sogenannter Optical Character Recognition (OCR) eignet sich besonders für sauber gedruckte Texte mit einer gewöhnlichen Schriftart und ohne manuell hinzugefügte Annotationen (vgl.

Annotation). Je mehr ein Text von diesem Ideal der Erfassbarkeit abweicht, desto schwieriger wird es, mittels automatisierter Verfahren ein annehmbares Ergebnis als Digitalisat zu erhalten. Die Tradition der optischen Zeichenerkennung reicht überraschend weit zurück. Der bereits in der Antike lebendige Traum, menschliche Fähigkeiten durch Maschinen nachzubilden, findet mit der Erfindung eines Retinascanners durch C. R. Carey 1870 in Massachusetts eine vorläufige Erfüllung (Eikvil 1993, 8).

Die Muster in heutigen OCR-Datenbanken, mit denen die erfassten Buchstaben abgeglichen werden, sind schriftartenspezifisch, weshalb auch die Wahl des OCR-Programms entscheidend zum Erfolg der Textdigitalisierung beiträgt. Die gängigsten Druckschriftarten sind in alle OCR-Softwares implementiert, moderne Anwendungen können zudem auf weitere Schriftarten „trainiert“ (vgl. Machine Learning) werden. Formatierungen wie eine Kursivierung oder Unterstreichung von Buchstaben oder Wörtern, unterschiedliche Schriftarten innerhalb eines Dokumentes und auch ein uneinheitlicher Buchstabenabstand (z. B. in Texten, die mit einer Schreibmaschine geschrieben wurden) können hier zu Schwierigkeiten führen. Ebenso können komplexere Layouts – z. B. ein Text mit mehreren Spalten oder Abbildungen – für die automatische Texterfassung große Herausforderungen bilden.

Fortgeschrittenere Programme gleichen die Ergebnisse in einem weiteren Schritt mit Wörterbucheinträgen ab und passen sie diesen bei Bedarf an. Problematisch sind hier historische Texte (die ein historisches Wörterbuch voraussetzen), uneinheitliche Schreibungen von Wörtern oder die Häufung von Eigennamen, die in der Regel nicht in einem Wörterbuch aufgeführt werden, jedoch häufig gerade in fiktionalen Texten vorkommen. Auch mehrsprachige Texte führen hier zu Schwierigkeiten. Auf Frakturschrift sind nur verhältnismäßig wenig Programme spezialisiert, am häufigsten wird hier der ABBYY FineReader (Schumacher 2024a) genannt, der jedoch nicht kostenfrei zu nutzen ist.

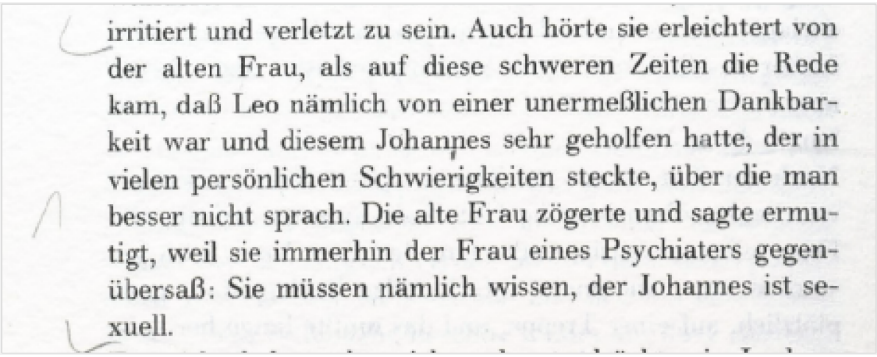
Die Angabe von Fehlerquoten bei der Evaluierung einzelner OCR-Programme ist mit Vorsicht zu genießen, da sie sich manchmal auf die Menge korrekt erfasster Wörter, manchmal aber auch auf die Menge korrekt erkannter Buchstaben in einem Textdokument bezieht. Zudem bleibt dabei der Effekt, den das jeweilige Preprocessing auf die Fehlerquote hat, im Dunkeln. Schließlich ist das Überprüfen von Fehlern in automatisch erfassten Texten sehr aufwändig und wird nur selten umfassend betrieben: eine Gegebenheit, die zu nicht repräsentativen Quoten führt. Das DFG-Projekt OCR-D widmet sich diesen Missständen und arbeitet außerdem daran, bestehende Verfahren zur automatischen Texterfassung in Hinblick auf die Besonderheiten von deutschsprachigen Texten des 16.–19. Jahrhunderts zu optimieren.

Komplizierter als die automatische Erkennung von gedruckter Schrift ist der Bereich der automatischen Handschriftenerkennung (vgl. HTR) (Digitale Manuskriptanalyse (Horstmann 2024a)). Handgeschriebene Schrift folgt sehr individuellen Mustern und variiert nicht nur zwischen unterschiedlichen Autor*innen stark, sondern häufig auch innerhalb des Werkes oder gar einzelner Texte von individuellen Autor*innen. Das Transkribus-Projekt (Horstmann 2024b) bietet im Umgang mit Handschriften und ihrer Digitalisierung eine verlässliche Anlaufstelle.

Wegen der genannten Schwierigkeiten greift man in der Regel gerade bei Handschriften auf die Methode des *Keying* (d. h. dem manuellen Abtippen des Textes) zurück, um eine wissenschaftlich verwertbare Genauigkeit des digitalisierten Textes zu erreichen. Exemplarische Untersuchungen der Textsammlung des Deutschen Textarchivs (Horstmann und Kern 2024) haben ergeben, dass die Erfassungsgenauigkeit sehr viel höher ist, wenn statt einer automatisierten optischen Zeichenerkennung das manuelle Verfahren des Double Keyings (vgl. Doublekeying) eingesetzt wird (Geyken u. a. 2012, 9). Beim Keying tippt lediglich eine Person den zu digitalisierenden Text händisch ab, wodurch es zu Lese- oder Flüchigkeitsfehlern kommen oder schlecht lesbarer Text falsch interpretiert werden kann. Im Double Keying-Verfahren hingegen erfassen zwei Personen den Text jeweils einmal manuell. Daraufhin werden Uneinheitlichkeiten in den beiden getrennt voneinander entstandenen Digitalisaten automatisch erfasst und von einer dritten Person händisch überprüft und ggf. korrigiert. Das Double Keying wird häufig an externe Anbieter ausgelagert, erfordert größere finanzielle Mittel als das OCR-Verfahren, überzeugt jedoch durch seine sehr hohe Genauigkeit (99.95% bis 99.98%). Auch bei diesen Quoten ist jedoch darauf zu achten, dass die Tests häufig nur beispielhaft anhand verhältnismäßig kleiner Textausschnitte durchgeführt wurden und es oft keine genaueren Angaben über die tatsächliche Textmenge, das Textgenre, die Art der gefundenen Fehler oder über die Transkriptoren gibt (Haaf, Wiegand und Geyken 2013).

5. Technische Grundlagen

Wenn Sie die Texte, die Sie erforschen wollen, selbst einscannen, achten Sie auf eine hohe Bildauflösung. Programme zur automatisierten Texterfassung kommen schnell an ihre Grenzen, wenn die zugrundegelegten Scans mangelhaft sind. An der folgenden Abbildung (ein Ausschnitt aus Ingeborg Bachmanns Erzählung *Das Gebell* (1972)) kann dies demonstriert werden.


irritiert und verletzt zu sein. Auch hörte sie erleichtert von der alten Frau, als auf diese schweren Zeiten die Rede kam, daß Leo nämlich von einer unermeßlichen Dankbarkeit war und diesem Johannes sehr geholfen hatte, der in vielen persönlichen Schwierigkeiten steckte, über die man besser nicht sprach. Die alte Frau zögerte und sagte ermutigt, weil sie immerhin der Frau eines Psychiaters gegenübersaß: Sie müssen nämlich wissen, der Johannes ist sexuell.

Scan eines Ausschnittes aus Ingeborg Bachmanns Erzählung Das Gebell *(1972)*

Die OCR-Funktion des Adobe Acrobat Pro erkennt darin folgenden Text:

LJrritiert und verletzt zu sein. Auch hörte sie erleichtert von
der alten Frau, als auf diese schweren Zeiten die Rede
kam, daß Leo nämlich von einer unermeßlichen Dankbar-.
keit war und diesem Johanpes sehr geholfen hatte, der in
vielen persönlichen Schwierigkeiten steckte, über die man
jl besser nicht sprach. Die alte Frau zögerte und sagte ermutigt,
weil sie immerhin der Frau eines Psychiaters gegenübersa.
: Sie müssen nämlich wissen, der Johannes ist seu
uell. . ” .

Wir sehen, dass das Programm offensichtlich Schwierigkeiten mit Schmutzflecken (z. B. „Johanpes“), Trennstrichen („seu uell“), gelegentlich dem Buchstaben ß „gegenübersa.:“ sowie manuellen Textannotationen („LJrritiert“) hat. Diese Fehler müssen nun entweder im Anschluss der automatischen Texterkennung manuell korrigiert werden, man beugt ihnen vor, indem man eine bessere Papier- und Scanqualität zugrunde legt, oder man findet ein OCR-Programm, das mit dem jeweiligen Text besser umgehen kann und besser verwertbare Ergebnisse liefert.

Eine OCR-Software identifiziert Abbildungen als solche und schließt sie aus der folgenden Texterkennung aus (Segmentierungsprozess). Ebenso werden Wortgrenzen, Zeilen und Absätze als solche festgestellt und Überschriften, Fußzeilen, Seitenzahlen etc. anhand ihres Abstandes zu den Blöcken des Haupttextes bzw. aufgrund divergierender Schriftarten gekennzeichnet. Jeder Punkt der Bilddatei wird entweder als Hintergrund oder als Text klassifiziert. In diesem Schritt der „Binarisierung“ (Rehbein 2017, 194) wurde der Fleck unter „Johannes“ als Text klassifiziert und führte daher zur fehlerhaften Texterkennung. Die als „Text“ identifizierten Pixel erhalten eine 1, die „Hintergrund“-Pixel eine 0. Das Schema der Einsen wird dann mit den Mustern in der OCR-Datenbank abgeglichen. Da dem Fleck unter dem „n“ des Johannes ebenfalls eine 1 zugeordnet wurde, entsprach das Schema eher dem Buchstaben und wurde als solcher umgesetzt.

Um Fehler zu vermeiden, können in einigen OCR- oder Scanprogrammen im Zuge des „Preprocessing“ (vgl. Preprocessing) diejenigen Bereiche des Dokumentes ausgewählt werden, die als Text erfasst werden sollen und die Abfolge der einzelnen Textfelder kann definiert werden (beispielsweise bei mehreren Spalten). Außerdem empfiehlt sich in diesem Schritt eine Optimierung der Farb- und Kontrastwerte (vgl. hierzu Kapitel 3 der DFG-Praxisregeln zur Digitalisierung).

Externe und weiterführende Links

- Preprocessing-Programme
  - ScanTailor (nicht für Mac): https://web.archive.org/save/http://scantailor.org/ (Letzter Zugriff: 04.06.2024)
  - ImageMagick: https://web.archive.org/save/https://www.imagemagick.org/script/index.php (Letzter Zugriff: 04.06.2024)
  - Unpaper: https://web.archive.org/save/https://github.com/Flameeyes/unpaper (Letzter Zugriff: 04.06.2024)
  - Leptonica: <https://web.archive.org/save/http://www.leptonica.com/ (Letzter Zugriff: 04.06.2024)
- OCR-Freeware
  - OCR4all: https://web.archive.org/save/https://lists.uni-wuerzburg.de/mailman/listinfo/ocr4all (Letzter Zugriff: 04.06.2024), (vgl. OCR4all (Schumacher 2024b))

  - Tesseract: https://web.archive.org/save/https://github.com/tesseract-ocr/tesseract (Letzter Zugriff: 04.06.2024)
  - OCRopus: https://web.archive.org/save/https://github.com/tmbdev/ocropy (Letzter Zugriff: 04.06.2024)
  - PDF XChange View (nicht für Mac): https://web.archive.org/save/https://www.tracker-software.com/product/pdf-xchange-viewer (Letzter Zugriff: 04.06.2024)
  - Google Docs: https://web.archive.org/save/https://support.google.com/drive/answer/176692?hl=de&co=GE-NIE.Platform=D...%5D (Letzter Zugriff: 04.06.2024)
- Kostenpflichtige OCR-Software
  - ABBYY FineReader: https://web.archive.org/save/https://www.abbyy.com/de-de/finereader/ (Letzter Zugriff: 04.06.2024), (vgl. Abbyy FineReader (Schumacher 2024a))
  - Adobe Reader Pro: https://web.archive.org/save/https://acrobat.adobe.com/de/de/acrobat/acrobat-pro.html (Letzter Zugriff: 04.06.2024)
  - Omnipage: https://web.archive.org/save/http://www.nuance.de/for-business/by-product/omnipage/index.html (Letzter Zugriff: 04.06.2024)
- DFG-Praxisregeln zur Digitalisierung: https://web.archive.org/save/https://web.archive.org/save/https://www.dfg.de/resource/blob/176108/898bf3574ad0ff3b1db525fa7d04c86c/12-151-v1216-de-data.pdf (Letzter Zugriff: 04.06.2024)
- OCR-Bibliothek aus dem Projekt OCR-D: https://web.archive.org/save/https://www.zotero.org/groups/418719/ocr-d (Letzter Zugriff: 04.06.2024)
- OCR-D: https://web.archive.org/save/http://ocr-d.de (Letzter Zugriff: 04.06.2024)
- Transkribus: https://web.archive.org/save/https://www.transkribus.org (Letzter Zugriff: 04.06.2024), (vgl. Transkribus (Horstmann 2024b))

Bibliographie

Berlin-Brandenburgische Akademie der Wissenschaften. 2018. Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. *Deutsches Textarchiv*. https://www.deutschestextarchiv.de/.

Deutsche Forschungsgemeinschaft und Digitalisierung. 2016. *DFG-Praxisregeln. „Digitalisierung“*. http://www.dfg.de/formulare/12_151/12_151_de.pdf (zugegriffen: 12. Juli 2018).

Eikvil, Line. 1993. *OCR. Optical Character Recognition*. https://www.nr.no/~eikvil/OCR.pdf (zugegriffen: 22. Januar 2018).

Geyken, Alexander, Mathhias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas und Frank Wiegand. 2012. TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. *Jahrbuch für Computerphilologie*. http://www.computerphilologie.de/jg09/geykenetal.html (zugegriffen: 22. Januar 2018).

Grubmüller, Klaus und Klaus Weimar. 1997. Edition. In: *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, hg. von Klaus Weimar, 1: A-G:414–418. Berlin, New York: de Gruyter.

Haaf, Susanne, Frank Wiegand und Alexander Geyken. 2013. Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text. *Journal of the Text Encoding Initiative*, Nr. 4. doi: 10.4000/jtei.739, (zugegriffen: 22. Januar 2018).

Horstmann, Jan. 2024a. Methodenbeitrag: Digitale Manuskriptanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3744, https://fortext.net/routinen/methoden/digitale-manuskriptanalyse.

———. 2024b. Toolbeitrag: Transkribus. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3746, https://fortext.net/tools/tools/transkribus.

Horstmann, Jan und Alexandra Kern. 2024. Ressourcenbeitrag: Deutsches Textarchiv (DTA). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3791, https://fortext.net/ressourcen/textsammlungen/deutsches-textarchiv-dta.

Kocher, Ursula. 2007. Textkritik. In: *Metzler Lexikon Literatur. Begriffe und Definitionen*, hg. von Dieter Burdorf, Christoph Fasbender, und Burkhard Moenninghoff, 761–762. Stuttgart, Weimar: Metzler.

Kraft, Herbert. 2001. *Editionsphilologie*. Frankfurt am Main (u.a.): Lang.

Nutt-Kofoth, Rüdiger. 2007. Textkritik und Textbearbeitung. In: *Handbuch Literaturwissenschaft*, hg. von Thomas Anz, 2: Methoden und Theorie:1–27. Stuttgart, Weimar: Metzler.

Rehbein, Malte. 2017. Digitalisierung. In: *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein, 179–198. Stuttgart: Metzler.

Rohr, Christian. 2015. *Historische Hilfswissenschaften: eine Einführung*. Wien (u.a.): Böhlau.

Schneider, Karin. 2003. Paläographie. In: *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, hg. von Jan-Dirk Müller, III: P-Z:1–3. Berlin, New York: de Gruyter.

———. 2014. *Paläographie und Handschriftenkunde für Germanisten: eine Einführung*. Berlin, New York: de Gruyter.

Schumacher, Mareike. 2024a. Toolbeitrag: Abbyy FineReader. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3742, https://fortext.net/tools/tools/abbyy-finereader.
———. 2024b. Toolbeitrag: OCR4all. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3743, https://fortext.net/tools/tools/ocr4all.
Sproll, Monika. 2007. Editionstechnik. In: *Metzler Lexikon Literatur. Begriffe und Definitionen*, hg. von Dieter Burdorf, Christoph Fasbender, und Burkhard Moenninghoff, 178. Stuttgart, Weimar: Metzler.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Double-keying Double-Keying ist eine Variante des Keying, bei der zwei Personen den Inhalt eines Dokumentes abtippen. Anschließend sucht ein Computerprogramm nach Differenzen zwischen den beiden Versionen. Gefundene Tippfehler werden dann von einer dritten Person korrigiert. So entstehen nahezu fehlerfreie Textdigitalisate.

HTR HTR steht für *Handwritten Text Recognition* und ist eine Form der Mustererkennung, wie auch die OCR. HTR bezeichnet die automatische Erkennung von Handschriften und die Umformung dieser in einen elektronischen Text. Die Automatisierung beruht auf einem Machine-Learning-Verfahren.

Keying In den Bibliotheks- und Textwissenschaften beschreibt Keying das manuelle Erfassen, also das Abtippen, eines Textes im Zuge seiner Digitalisierung (siehe auch Double-Keying).

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.